A simple hierarchical Gaussian process model for housing prices in Finland

Ville Mäkinen

13.3.2021

Introduction

This document describes a simple hierarchical Gaussian process model for housing prices in Finland. The objective of the model is provide a working example of a Gaussian process in the housing prices context in contrast to the previous work. Moreover, a new, more detailed, data set was collected such that the group-level model is defined over all post code areas in Finland. The presented Gaussian process model now incorporates the geographical effects on the basis of the post code areas.

Any questions or comments can be sent to ville piste ka piste makinen ät gmail piste com.

Data

The data was collected via web scraping from the *asuntojen.hintatieto.fi*-service of the Finnish Ministery of the Environment and the Housing Finance and Development Centre of Finland.

Data usage

Data was split into three data sets; estimation set, 'in-sample' test set and 'out-of-sample' test set. The splits were done by separating 5 % of the post codes found in the raw data to form the out-of-sample test set. Next, out of the remaining data, 20 % of the available observations are set aside as the in-sample test set. The remaining observations are used as the estimation set. The motivation for separating a out-of-sample test set is to check whether the Gaussian process captures geographical properties of the data.

Variable preprocessing

For the regression models, each explanatory variable is centered and scaled. Centering is done subtracting the estimation set means from the untreated variables. Scaling is done by dividing the centered variables by 2 standard deviations calculated from the estimation set.

The response variable $\log \mathrm{Price}_i^*$ is calculated by taking the natural logarithm for the recorded sales prices and then subtracting the estimation set mean from the variable, i.e.

$$\log \mathrm{Price}_i^* = \log \mathrm{Price}_i - \overline{\log \mathrm{Price}}_{\text{Estimation set}} \tag{1}$$

where the term $\overline{\log \mathrm{Price}}_{\text{Estimation set}}$ is the mean log price calculated from the estimation set. The price predictions for the in-sample and out-of-sample test set prices are generated by first predicting the (centered) log prices for each observation, then adding the estimation set mean log price from the estimation set.

Models

Two models were estimated using the newly collected data. Since the response variable are centered, no (explicit) intercept terms are included.

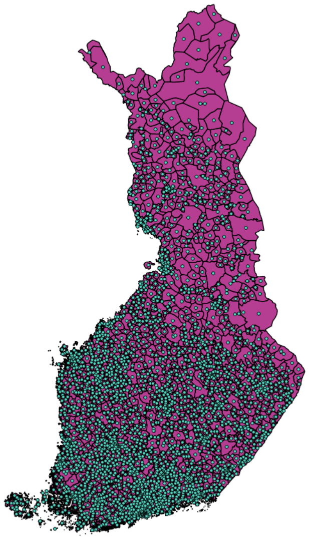

Figure 1: Post code areas

Simple linear hierarchical model

Structure

The likelihood of the model is given as

$$\log \text{Price}_i^* \sim \text{N}(\mu_i, \sigma^2)$$

where the expected value μ_i is determined by the sum

$$\begin{aligned}
\mu_i \quad = \quad & \beta_{j[i]}^{\text{PostCode}} + \\
& \beta_{\text{Sqm}} \text{SqmStandardized}_i + \\
& \beta_{\text{OwnPropertyDummy}} \text{OwnPropertyDummyStandardized}_i + \\
& \beta_{\text{RowHouseDummy}} \text{RowHouseDummyStandardized}_i + \\
& \beta_{\text{TownHouseDummy}} \text{TownHouseDummyStandardized}_i + \\
& \beta_{\text{SaunaDummy}} \text{SaunaDummyStandardized}_i + \\
& \beta_{\text{ConditionUnrecordedDummy}} \text{ConditionUnrecordedDummyStandardized}_i + \\
& \beta_{\text{ConditionGoodDummy}} \text{ConditionGoodDummyStandardized}_i + \\
& \beta_{\text{ConditionAdequateDummy}} \text{ConditionAdequateDummyStandardized}_i + \\
& \beta_{\text{AgeOfBuilding}} \text{AgeOfBuildingStandardized}_i
\end{aligned}$$

where the term $\beta_j^{\text{PostCode}}$ is the group-specific coefficient defined for each post code. The group-level model parameter priors are the following:

$$\begin{aligned}
\beta_j^{\text{PostCode}} \quad & \sim \quad \text{N}(0, \sigma_{\text{PostCode}}^2), \\
\sigma_{\text{PostCode}} \quad & \sim \quad \text{Half-Cauchy}(0, 1).
\end{aligned}$$

The prior distribution choices for the coefficients are

$$
\begin{aligned}
\beta_{\text{Sqm}} &\sim \mathrm{N}(0,1),\\
\beta_{\text{OwnPropertyDummy}} &\sim \mathrm{N}(0,1),\\
\beta_{\text{RowHouseDummy}} &\sim \mathrm{N}(0,1),\\
\beta_{\text{TownHouseDummy}} &\sim \mathrm{N}(0,1),\\
\beta_{\text{SaunaDummy}} &\sim \mathrm{N}(0,1),\\
\beta_{\text{ConditionUnrecordedDummy}} &\sim \mathrm{N}(0,1),\\
\beta_{\text{ConditionGoodDummy}} &\sim \mathrm{N}(0,1),\\
\beta_{\text{ConditionAdequateDummy}} &\sim \mathrm{N}(0,1),\\
\beta_{\text{AgeOfBuilding}} &\sim \mathrm{N}(0,1)
\end{aligned}
$$

and the prior distribution for the residual variance parameter is

$$\sigma \sim \text{Half-Cauchy}(0,1). \tag{2}$$

Estimates

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
β_{Sqm}	0.53	0.00	0.01	0.52	0.52	0.53	0.53	0.54	14020.54	1.00
$\beta_{\text{OwnPropertyDummy}}$	0.05	0.00	0.01	0.04	0.05	0.05	0.06	0.07	12912.76	1.00
$\beta_{\text{RowHouseDummy}}$	0.17	0.00	0.01	0.16	0.17	0.17	0.17	0.18	10617.36	1.00
$\beta_{\text{TownHouseDummy}}$	0.24	0.00	0.01	0.22	0.23	0.24	0.24	0.25	9171.00	1.00
$\beta_{\text{SaunaDummy}}$	0.12	0.00	0.01	0.10	0.11	0.12	0.12	0.13	18711.42	1.00
$\beta_{\text{ConditionUnrecordedDummy}}$	0.24	0.00	0.01	0.22	0.23	0.24	0.25	0.26	8229.70	1.00
$\beta_{\text{ConditionGoodDummy}}$	0.51	0.00	0.02	0.48	0.50	0.51	0.52	0.55	7862.10	1.00
$\beta_{\text{ConditionAdequateDummy}}$	0.25	0.00	0.01	0.22	0.24	0.24	0.25	0.27	8115.02	1.00
$\beta_{\text{AgeOfBuilding}}$	-0.33	0.00	0.01	-0.34	-0.33	-0.33	-0.32	-0.31	16061.16	1.00
σ	0.30	0.00	0.00	0.30	0.30	0.30	0.30	0.31	24067.47	1.00
σ_{PostCode}	0.67	0.00	0.02	0.63	0.65	0.66	0.68	0.70	404.78	1.00

Table 1: Simple linear model coefficient estimates

The below figure plots the posterior distributions of the post code effects for post codes with the 30 highest and lowest mean effects under the simple linear hierarchical model.

Gaussian process model

Structure

The likelihood of the model is given as

$$\log \text{Price}_i^* \sim \mathrm{N}(\mu_i, \sigma^2)$$

where the expected value μ_i is determined by the sum

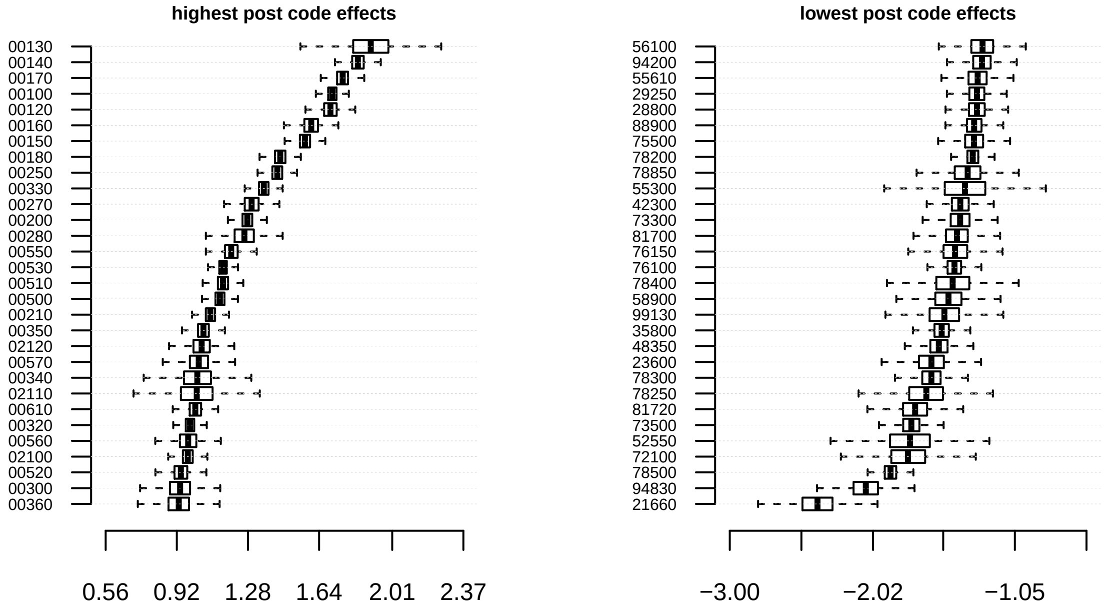

Figure 2: Posteriors for post code intercepts, simple linear hierarchical model

$$
\begin{aligned}
\mu_i \quad = \quad & \beta^{\text{PostCode}}_{j[i]} + \\
& \beta_{\text{Sqm}}\text{SqmStandardized}_i + \\
& \beta_{\text{OwnPropertyDummy}}\text{OwnPropertyDummyStandardized}_i + \\
& \beta_{\text{RowHouseDummy}}\text{RowHouseDummyStandardized}_i + \\
& \beta_{\text{TownHouseDummy}}\text{TownHouseDummyStandardized}_i + \\
& \beta_{\text{SaunaDummy}}\text{SaunaDummyStandardized}_i + \\
& \beta_{\text{ConditionUnrecordedDummy}}\text{ConditionUnrecordedDummyStandardized}_i + \\
& \beta_{\text{ConditionGoodDummy}}\text{ConditionGoodDummyStandardized}_i + \\
& \beta_{\text{ConditionAdequateDummy}}\text{ConditionAdequateDummyStandardized}_i + \\
& \beta_{\text{AgeOfBuilding}}\text{AgeOfBuildingStandardized}_i
\end{aligned}
$$

where the term $\beta^{\text{PostCode}}_{j}$ is the group-specific coefficient defined for each post code. The group-specific coefficient are determined through a Gaussian process model such that

$$\beta^{\text{PostCode}}_{j} \sim \text{Multivariate-Normal}(0, \Sigma)$$

where the covariance matrix Σ is defined with the kernel

$$K_{ij} = \alpha^2 \exp\left(-\frac{1}{2\rho^2}(\text{Distance between post code centroids i and j})^2\right) + \sigma^2_{\text{GP}}\delta_{ij}$$

where the terms α, ρ and σ^2_{GP} are parameters and the term δ_{ij} is the Kronecker delta. The parameters have the following prior distributions:

$$
\begin{aligned}
\alpha &\sim \text{Half-Cauchy}(0, 1), \\
\rho &\sim \text{GeneralizedInverseGaussian}(\text{p} = 1,\ \text{a} = 1,\ \text{b} = 1)
\end{aligned}
$$

and

$$\sigma_{\text{GP}} \sim \text{Half-Cauchy}(0, 1).$$

Finally, the priors for the coefficients are

$$\begin{aligned}
\beta_{\text{Sqm}} &\sim \text{N}(0, 1), \\
\beta_{\text{OwnPropertyDummy}} &\sim \text{N}(0, 1), \\
\beta_{\text{RowHouseDummy}} &\sim \text{N}(0, 1), \\
\beta_{\text{TownHouseDummy}} &\sim \text{N}(0, 1), \\
\beta_{\text{SaunaDummy}} &\sim \text{N}(0, 1), \\
\beta_{\text{ConditionUnrecordedDummy}} &\sim \text{N}(0, 1), \\
\beta_{\text{ConditionGoodDummy}} &\sim \text{N}(0, 1), \\
\beta_{\text{ConditionAdequateDummy}} &\sim \text{N}(0, 1), \\
\beta_{\text{AgeOfBuilding}} &\sim \text{N}(0, 1)
\end{aligned}$$

and the prior for the residual variance is

$$\sigma \sim \text{Half-Cauchy}(0, 1).$$

Estimates

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
β_{Sqm}	0.53	0.00	0.01	0.51	0.52	0.53	0.53	0.54	4046.05	1.00
$\beta_{\text{OwnPropertyDummy}}$	0.06	0.00	0.01	0.05	0.05	0.06	0.06	0.07	5104.99	1.00
$\beta_{\text{RowHouseDummy}}$	0.17	0.00	0.01	0.16	0.16	0.17	0.17	0.18	3800.83	1.00
$\beta_{\text{TownHouseDummy}}$	0.24	0.00	0.01	0.23	0.24	0.24	0.25	0.26	3143.38	1.00
$\beta_{\text{SaunaDummy}}$	0.12	0.00	0.01	0.11	0.11	0.12	0.12	0.13	5953.97	1.00
$\beta_{\text{ConditionUnrecordedDummy}}$	0.25	0.00	0.01	0.22	0.24	0.25	0.25	0.27	2248.21	1.00
$\beta_{\text{ConditionGoodDummy}}$	0.52	0.00	0.02	0.49	0.51	0.52	0.53	0.55	2228.87	1.00
$\beta_{\text{ConditionAdequateDummy}}$	0.25	0.00	0.01	0.22	0.24	0.25	0.26	0.28	2301.32	1.00
$\beta_{\text{AgeOfBuilding}}$	-0.32	0.00	0.01	-0.33	-0.33	-0.32	-0.32	-0.31	5426.88	1.00
σ	0.30	0.00	0.00	0.30	0.30	0.30	0.30	0.31	7456.27	1.00

Table 2: Coefficient estimates

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
ρ	23.79	0.09	1.30	21.24	22.89	23.80	24.67	26.34	224.05	1.01
α	0.77	0.00	0.05	0.67	0.73	0.76	0.80	0.88	239.72	1.01
σ_{GP}	0.20	0.00	0.01	0.19	0.20	0.20	0.21	0.22	508.41	1.00

Table 3: GP model parameter estimates

The below figure plots the posterior distributions of the post code effects for post codes with the 30 highest and lowest mean effects under the Gaussian process model.

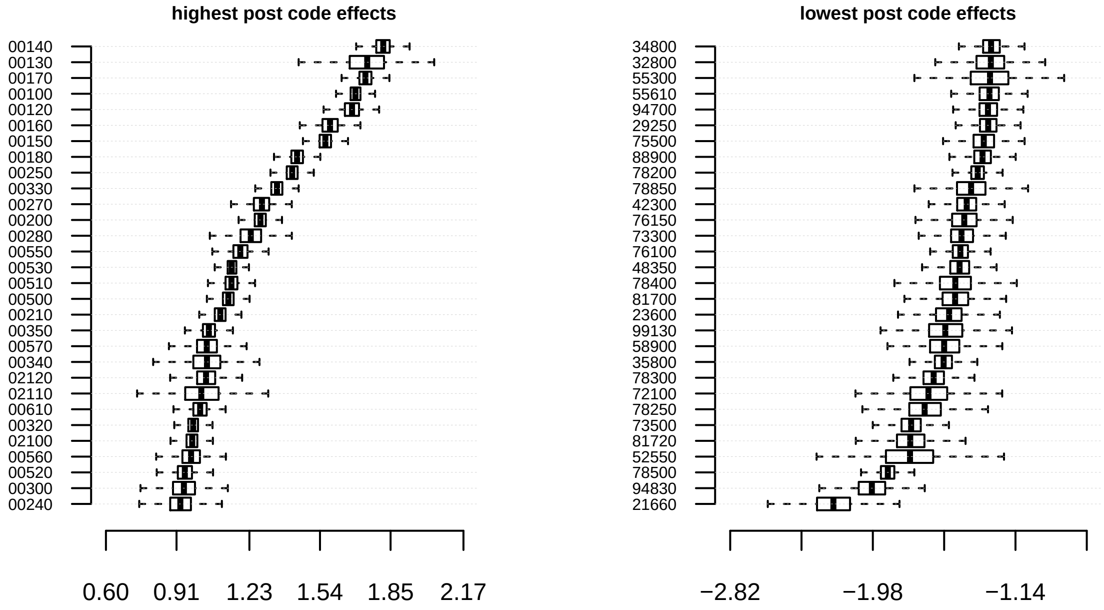

Figure 3: Posteriors for post code intercepts, Gaussian process model

Results

LOO comparison

The loo-statistics printout for the Gaussian process model is the following:

```
##
## Computed from 4000 by 18267 log-likelihood matrix
##
##          Estimate    SE
## elpd_loo  -4522.9 425.6
## p_loo       841.7  62.0
## looic      9045.9 851.2
## ------
## Monte Carlo SE of elpd_loo is NA.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     18173 99.5%   273
##  (0.5, 0.7]   (ok)          76  0.4%   75
##    (0.7, 1]   (bad)         12  0.1%   14
##    (1, Inf)   (very bad)     6  0.0%   3
## See help('pareto-k-diagnostic') for details.
```

The loo-statistics printout for the simple linear hierarchical model is the following:

```
##
## Computed from 12000 by 18267 log-likelihood matrix
##
##          Estimate    SE
## elpd_loo  -4599.1 426.6
## p_loo       938.5  60.9
## looic      9198.2 853.1
## ------
## Monte Carlo SE of elpd_loo is NA.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     18100 99.1%   523
##  (0.5, 0.7]   (ok)         126  0.7%   120
##    (0.7, 1]   (bad)         35  0.2%   22
##    (1, Inf)   (very bad)     6  0.0%   5
## See help('pareto-k-diagnostic') for details.
```

Both loo-statistics printouts indicate problems with the models: Ideally, there shouldn't observations with pareto k greater than 0.7 or these should be handled with exact LOO calculations. The some of these problematic observations are seemingly outliers with likely data entry issues, these include e.g. observations with price per square meters of 74 e/m^2 or buildings built in the year 1056. Perhaps the results could be (somewhat) improved by using a robust likelihood distribution.

Model comparison figures indicate that the Gaussian process model outperforms the simple linear hierarchical model:

```r
loo_compare(loo_fit_gp, loo_fit_simple_lin_reg)
```

```
##        elpd_diff se_diff
## model1   0.0       0.0
## model2 -76.2      16.8
```

Predictive distributions

Calibration

The PIT histograms indicate that neither model is well-calibrated - ideally the PIT histogram should be match the density of Uniform(0,1).

Point predictions

Point predictions are generated by taking the means from the predictive distribution samples.

Point predictions for in-sample test set

From the figure below it can be seen that the point predictions are essentially the same for both models for the test set with the in-sample post codes.

Point predictions for out-of-sample test set

For the out-of-sample post codes, the point predictions of the Gaussian process model match the observed prices somewhat better than the point predictions from the simple linear hierarchical model. Thus the Gaussian process approach captures at least some of the spatial effects of the true data generating process.

The graphs below depict the point predictions for specific post codes. The graphs show that there are post codes where the Gaussian process approach is necessary (e.g. 00260, 02330, 88610) as well as post codes where the Gaussian process produces worse predictions (e.g. 20720).

Shiny app for price predictions for the GP model

https://ville-makinen.shinyapps.io/gp_predictions_shiny_app/

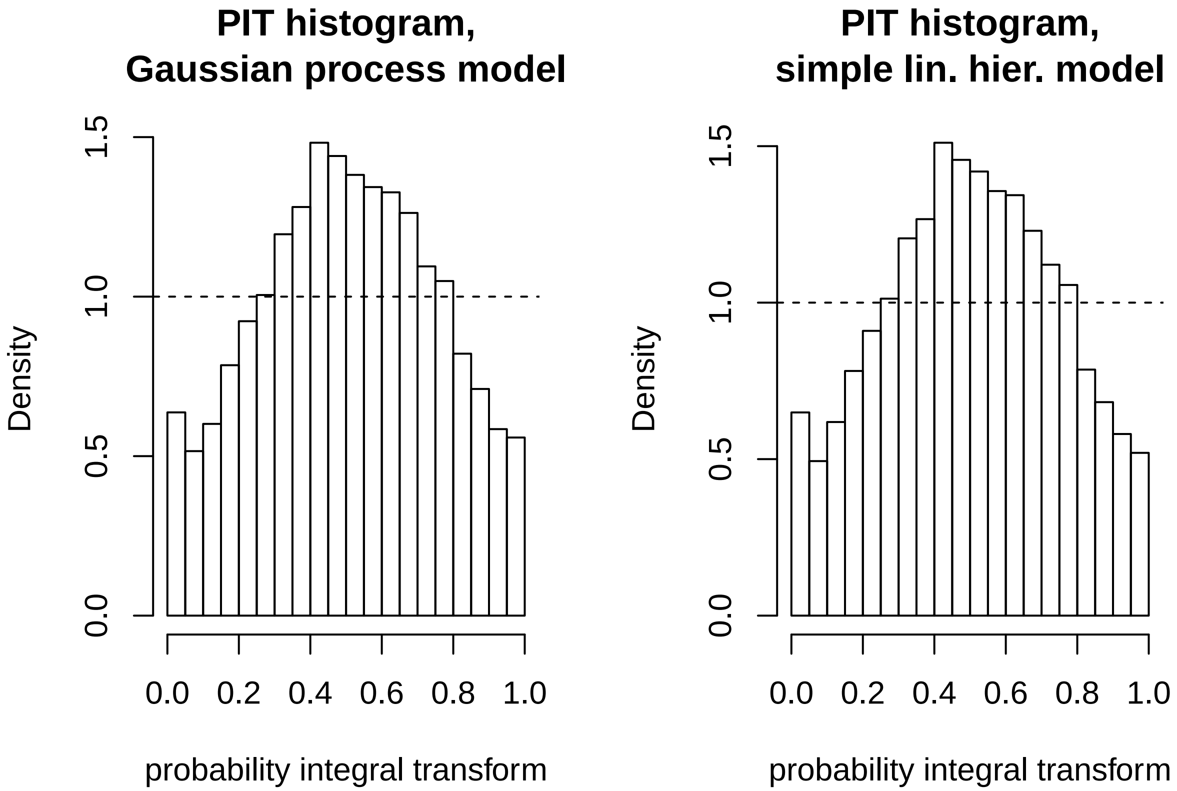

Figure 4: PIT histograms

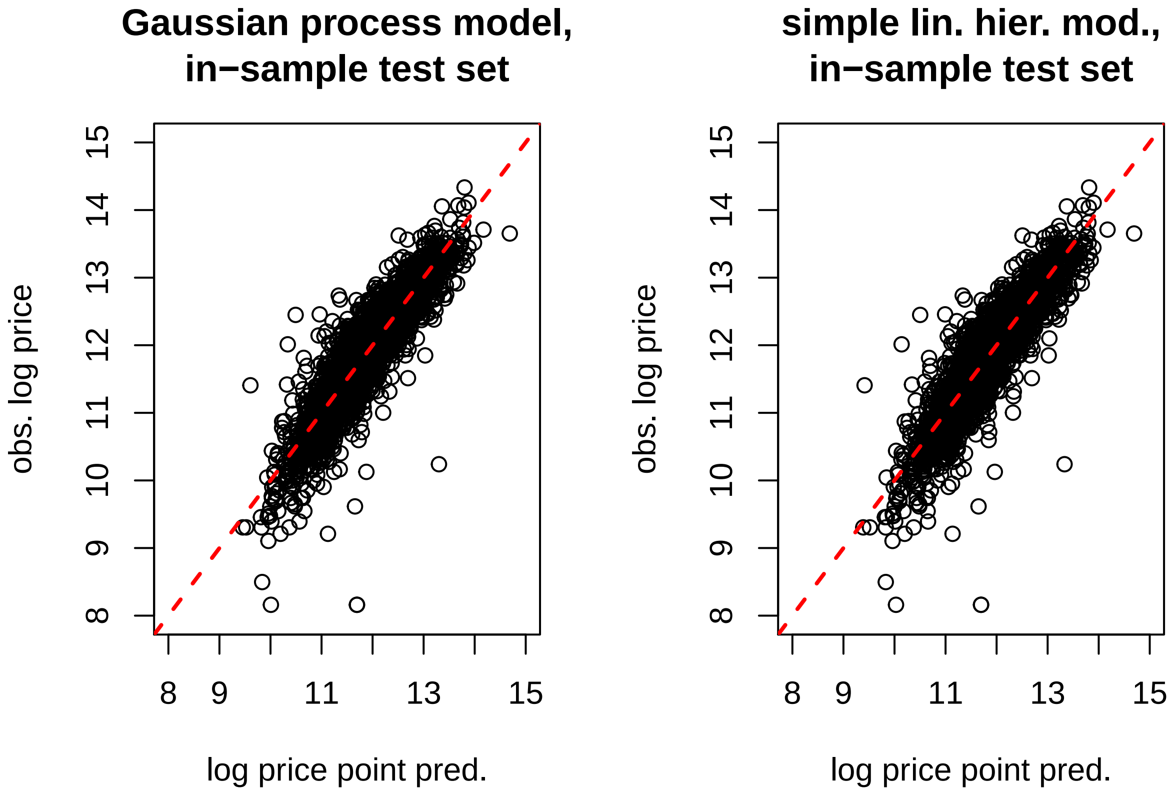

Figure 5: In-sample test set point predictions

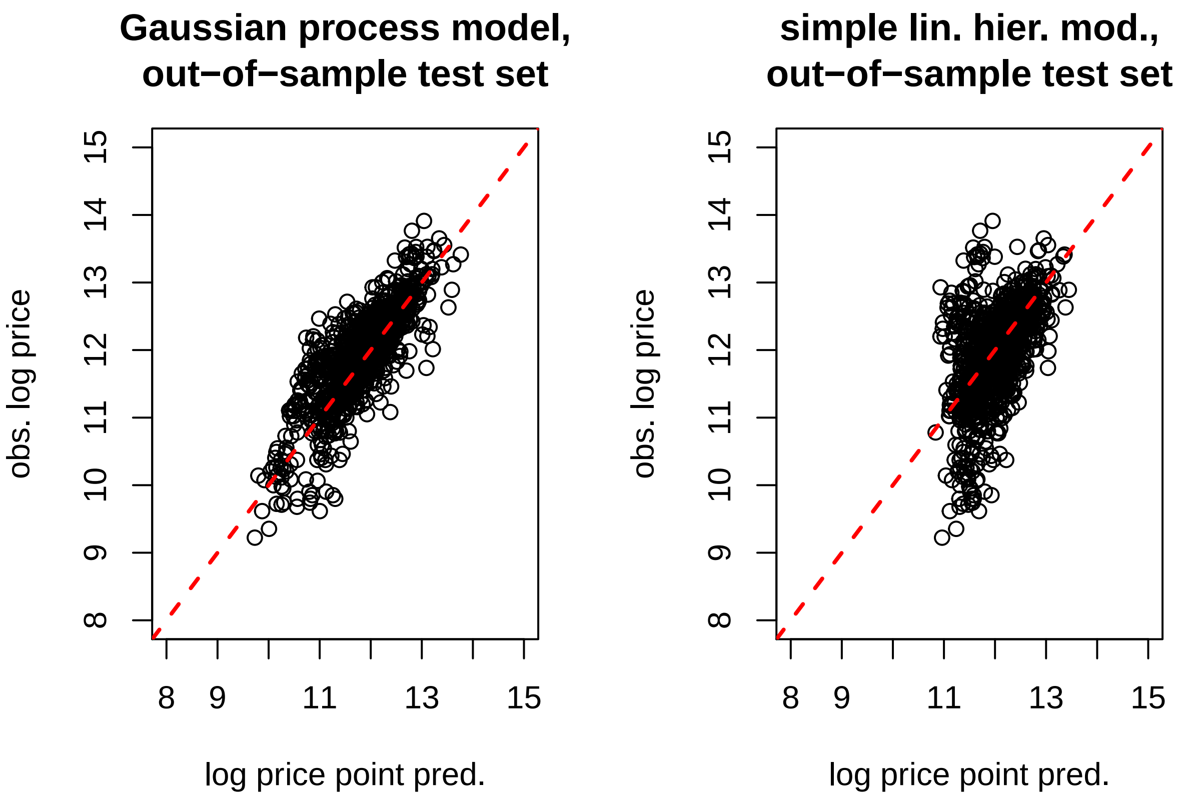

Figure 6: Out-of-sample test set point predictions

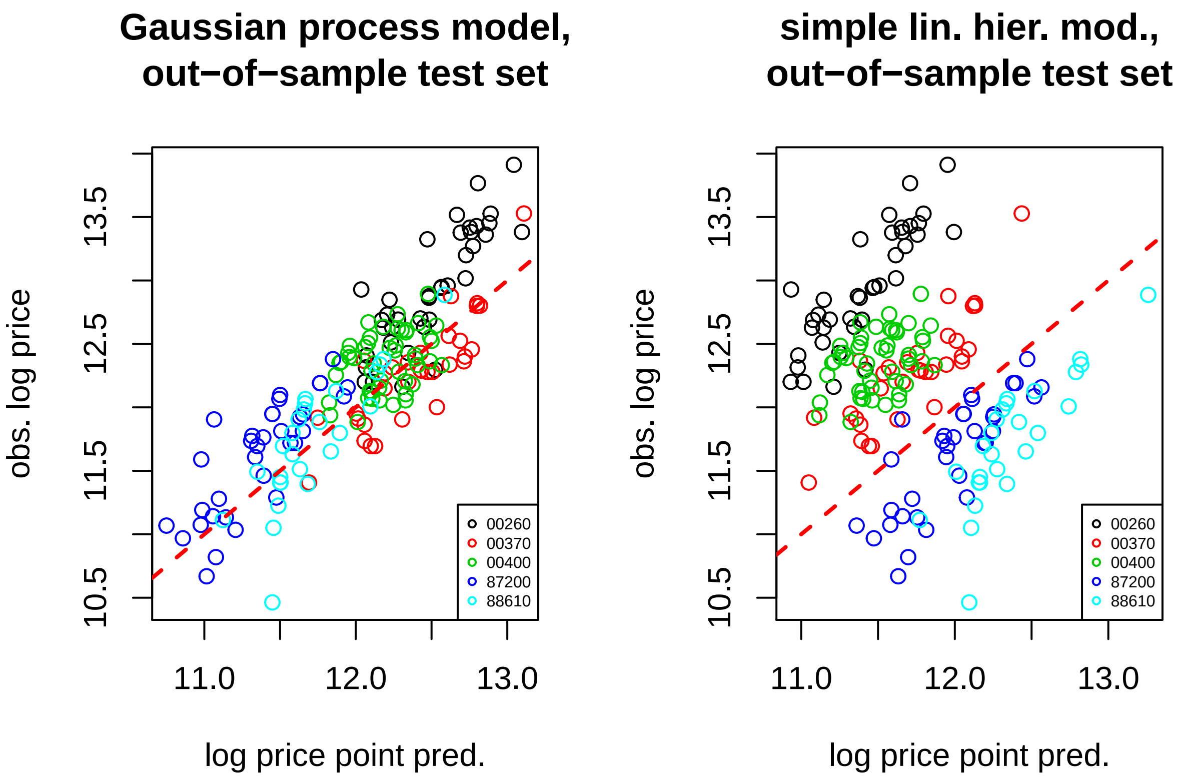

Figure 7: Out-of-sample test set point predictions, contd.

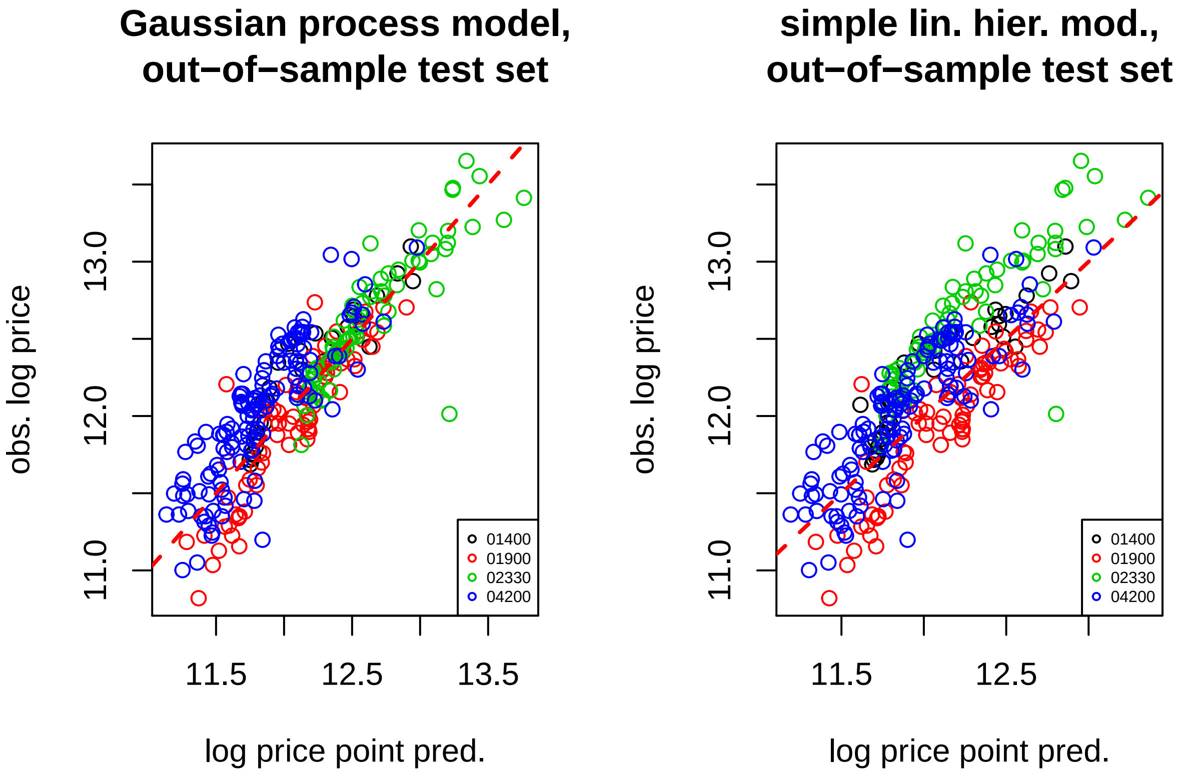

Figure 8: Out-of-sample test set point predictions, contd.

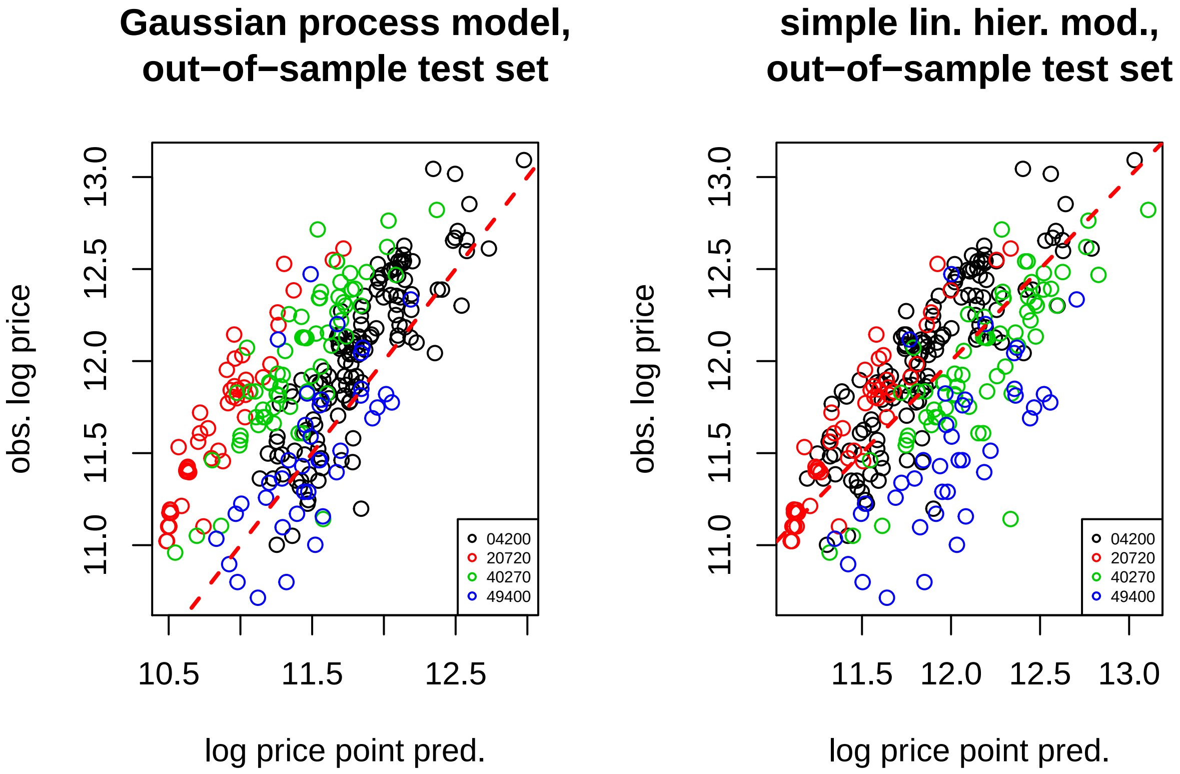

Figure 9: Out-of-sample test set point predictions, contd.